

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-282321

(43)Date of publication of application : 31.10.1997

(51)Int.Cl. G06F 17/28
G06F 17/27
G10L 3/00

JP 09-282321
U.S. PTO
09/745795

(21)Application number : 08-198950 (71)Applicant : ATR ONSEI HONYAKU TSUSHIN
KENKYUSHO:KK

(22)Date of filing : 29.07.1996 (72)Inventor : SHIODA AKIRA
IIDA HITOSHI

(30)Priority

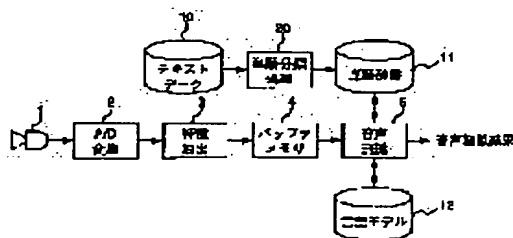
Priority number : 08 27809 Priority date : 15.02.1996 Priority country : JP

(54) WORD CLASSIFICATION PROCESSING METHOD AND DEVICE THEREFOR, AND VOICE RECOGNIZER

(57)Abstract:

PROBLEM TO BE SOLVED: To obtain a word classification result having a well-balanced hierarchical structure by classifying plural words into plural classes in the form of a binary tree having hierachized lower, intermediate and upper layers.
SOLUTION: The word classification processing part 20 classifies the words included in the text data stored in a text data memory 10 by assigning the words of comparatively low appearance frequency and the words of high rates to be adjacent to the same word in the same classes respectively. Then, the part 20 classes the word classification result into the intermediate, upper and lower layers. Then, the words are classified in order of intermediate, upper and lower layers and based on the prescribed average mutual information content, i.e., a global (overall) cost function set for all words included in the text data.

The classified words are stored in a word dictionary memory 11 in the form of a word dictionary. In such word classification processing, it is possible to obtain the word classification result that has a well-balanced hierarchical structure and also is globally optimized.



LEGAL STATUS

[Date of request for examination] 29.07.1996

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]	3043625
[Date of registration]	10.03.2000
[Number of appeal against examiner's decision of rejection]	
[Date of requesting appeal against examiner's decision of rejection]	
[Date of extinction of right]	

Copyright (C); 1998,2000 Japanese Patent Office

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平9-282321

(43) 公開日 平成9年(1997)10月31日

(51) Int.Cl. ⁶	識別記号	庁内整理番号	F I	技術表示箇所
G 0 6 F 17/28			G 0 6 F 15/38	C
17/27			G 1 0 L 3/00	5 6 1 G
G 1 0 L 3/00	5 6 1		G 0 6 F 15/38	E

審査請求 有 請求項の数 5 O L (全 22 頁)

(21) 出願番号 特願平8-198950

(22) 出願日 平成8年(1996)7月29日

(31) 優先権主張番号 特願平8-27809

(32) 優先日 平8(1996)2月15日

(33) 優先権主張国 日本 (J P)

(71) 出願人 593118597

株式会社エイ・ティ・アール音声翻訳通信
研究所

京都府相楽郡精華町大字乾谷小字三平谷 5
番地

(72) 発明者 潮田 明

京都府相楽郡精華町大字乾谷小字三平谷 5
番地 株式会社エイ・ティ・アール音声翻
訳通信研究所内

(72) 発明者 飯田 仁

京都府相楽郡精華町大字乾谷小字三平谷 5
番地 株式会社エイ・ティ・アール音声翻
訳通信研究所内

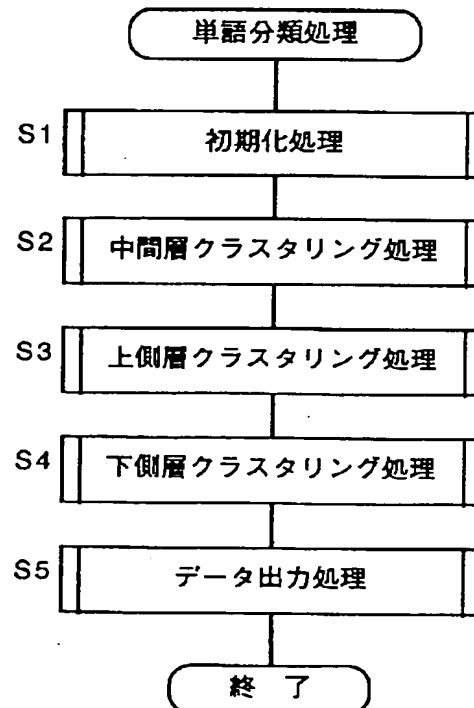
(74) 代理人 弁理士 青山 葆 (外 2 名)

(54) 【発明の名称】 単語分類処理方法、単語分類処理装置及び音声認識装置

(57) 【要約】

【課題】 単語分類処理によりバランスのとれた階層構造を有しかつ全体的に最適化された単語分類結果を得ることができる単語分類処理方法、単語分類処理装置、及びその単語分類処理装置を備えた音声認識装置を提供する。

【解決手段】 複数の単語を含むテキストデータに対して、互いに異なるすべての複数 v 個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数 v 個のクラスに割り当て、上記複数 v 個のクラスの単語のうち出現頻度が高い v 個未満の $(c + 1)$ 個のクラスの単語を 1 つのウィンドウ内のクラスの単語として第 1 のメモリに記憶し、当該クラスの単語に基づいて、所定の平均相互情報量が最大となるように、複数の単語を二分木の形式で複数 c 個のクラスに分類して単語分類結果を表わす全体のツリー図の中間層を求め、当該中間層を中心として、上側層と、中間層の各クラス毎の複数の下側層とを求めて、全体のツリー図を求める。



【特許請求の範囲】

【請求項1】 複数の単語を含むテキストデータに対して、互いに異なるすべての複数の v 個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数の v 個のクラスに割り当てるステップと、

上記複数の v 個のクラスの単語のうち出現頻度が高い v 個未満の $(c+1)$ 個のクラスの単語を1つのウィンドウ内のクラスの単語として第1の記憶装置に記憶するステップと、

上記第1の記憶装置に記憶された1つのウィンドウ内のクラスの単語に基づいて、互いに異なる第1のクラスの単語と第2のクラスの単語とが隣接して出現する確率を、上記第1のクラスの単語の出現確率と第2のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数の c 個のクラスに分類し、分類された複数の c 個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数の c 個のクラスとして第2の記憶装置に記憶するステップと、

上記第2の記憶装置に記憶された中間層の複数の c 個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第3の記憶装置に記憶するステップと、

上記第2の記憶装置に記憶された中間層の複数の c 個のクラスの各クラス毎に、上記中間層の複数の c 個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第4の記憶装置に記憶するステップと、

上記第4の記憶装置に記憶された上記ツリー図の下側層を、上記第2の記憶装置に記憶された上記中間層の複数の c 個のクラスと連結する一方、上記第3の記憶装置に記憶された上記ツリー図の上側層を、上記第2の記憶装置に記憶された上記中間層の複数の c 個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第5の記憶装置に記憶するステップとを備えたことを特徴とする単語分類処理方法。

【請求項2】 上記分類された複数の c 個のクラスを上記第2の記憶装置に記憶するステップは、上記第1の記憶装置に記憶された1つのウィンドウよりも外側のクラスが存在し、又は上記1つのウィンドウ内のクラスが c 個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする請求項1記載の単語分類処理方法。

【請求項3】 複数の単語を含むテキストデータに対し

て、互いに異なるすべての複数の v 個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数の v 個のクラスに割り当てる第1の制御手段と、

上記複数の v 個のクラスの単語のうち出現頻度が高い v 個未満の $(c+1)$ 個のクラスの単語を1つのウィンドウ内のクラスの単語として第1の記憶装置に記憶する第2の制御手段と、

上記第1の記憶装置に記憶された1つのウィンドウ内のクラスの単語に基づいて、互いに異なる第1のクラスの単語と第2のクラスの単語とが隣接して出現する確率を、上記第1のクラスの単語の出現確率と第2のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数の c 個のクラスに分類し、分類された複数の c 個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数の c 個のクラスとして第2の記憶装置に記憶する第3の制御手段と、

上記第2の記憶装置に記憶された中間層の複数の c 個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第3の記憶装置に記憶する第4の制御手段と、

上記第2の記憶装置に記憶された中間層の複数の c 個のクラスの各クラス毎に、上記中間層の複数の c 個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第4の記憶装置に記憶する第5の制御手段と、

上記第4の記憶装置に記憶された上記ツリー図の下側層を、上記第2の記憶装置に記憶された上記中間層の複数の c 個のクラスと連結する一方、上記第3の記憶装置に記憶された上記ツリー図の上側層を、上記第2の記憶装置に記憶された上記中間層の複数の c 個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第5の記憶装置に記憶する第6の制御手段とを備えたことを特徴とする単語分類処理装置。

【請求項4】 上記第3の制御手段は、上記第1の記憶装置に記憶された1つのウィンドウよりも外側のクラスが存在し、又は上記1つのウィンドウ内のクラスが c 個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする請求項3記載の単語分類処理装置。

【請求項5】 入力される発声音の音声信号に基づいて、請求項3又は4記載の単語分類処理装置によって複数の単語が複数のクラスに分類された単語分類結果を含む単語辞書と、所定の隠れマルコフモデルとを参照して

上記発声音声を音声認識する音声認識手段を備えたことを特徴とする音声認識装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、音声認識装置、形態素解析装置、及び構文解析装置のための単語分類処理方法及び単語分類処理装置、並びに、上記単語分類処理装置を備えた音声認識装置に関する。

【0002】

【従来の技術】単語の分類体系は、音声認識装置、形態素解析装置や構文解析装置において処理を円滑に行う上で非常に重要な知識の1つである。この単語の分類体系を構築するための1つの方法として、大量のテキストデータに基づいて単語間の相互情報量を用いた方法（以下、第1の従来例という。）が、例えば、従来技術文献「Peter Brown, et al, "Class-Based n-gram Models of Natural Language", Computational Linguistics, Vol. 18, No. 4, pp. 467-479, 1992年」において提案されている。この従来例の方法においては、n-グラムモデルを用いて所定の相互情報量を計算して英単語の分類を行っている。

【0003】しかしながら、第1の従来例の相互情報量による分類処理方法を用いて英単語を分類した場合、出現頻度の低い単語が不適切に分類される場合が多いという問題点があった。この問題点を解決するために、単語分類処理装置及び音声認識装置（以下、第2の従来例という。）が、本出願人により特願平7-056918号の特許出願において提案されている。

【0004】当該第2の従来例の単語分類処理装置は、単語のn-グラムを利用して、同一の単語に隣接する割合の多い単語を同一のクラスに割り当てるという基準で複数の単語を複数のクラスに分類する第1の分類手段と、上記第1の分類手段によって分類された複数の単語に対して、すべての単語の出現頻度を調べ、互いに異なる第1のクラスの単語と第2のクラスの単語とが隣接して出現する頻度を、上記第1のクラスの単語の出現頻度と第2のクラスの単語の出現頻度との積に対する相対的な頻度の割合を表わす所定の相互情報量が最大となるように、上記複数の単語を二分木の形式で複数のクラスに分類する第2の分類手段とを備えたことを特徴としている。ここで、上記第2の分類手段は、好ましくは、上記第1の分類手段によって分類された複数の単語に対して、すべての単語の出現頻度を調べ、出現頻度の高い単語から順に、所定の複数N個のクラスに割り当て、N個のクラスの中で上記相互情報量が最大である2つのクラスを1つのクラスとしてまとめることにより、(N-1)個のクラスに分類し、クラスに割り当てられていない単語の中で、出現頻度が最大のものを新たにN番目のクラスとして割り当て、すべての単語がN個のクラスに割り当てられるまで、上記の処理を繰り返し、現在ある

クラスから上記相互情報量が最大である2つのクラスを1つのクラスとしてまとめ、この処理を1個のクラスにまとまるまで繰り返す。これにより、単語分類をより安定に実行することができ、テキストデータから単語の分類体系を自動的に獲得するときに、より精密で正確な分類体系を得ることができるという特徴を有する。

【0005】

【発明が解決しようとする課題】上記第1の従来例の相互情報量による分類処理方法を用いて英単語を分類した場合、出現頻度の低い単語が不適切に分類される場合が多い。この原因としては、分離結果がバランスのとれた階層構造となっていないためであると考えられる。

【0006】また、上記第2の従来例においては、互いに異なる第1のクラスの単語と第2のクラスの単語とが隣接して出現する頻度を、上記第1のクラスの単語の出現頻度と第2のクラスの単語の出現頻度との積に対する相対的な頻度の割合を表わす所定の相互情報量が最大となるように、上記複数の単語を二分木の形式で複数のクラスに分類しているの、上記第1のクラスの単語と上記第2のクラスの単語においては、局所的に最適化された単語分類結果を得ることができるが、全体的に最適化された単語分類結果を得ることができないという問題点があった。

【0007】本発明の目的は以上の問題点を解決し、単語分類処理によりバランスのとれた階層構造を有しかつ全体的に最適化された単語分類結果を得ることができる単語分類処理方法、単語分類処理装置、及びその単語分類処理装置を備えた音声認識装置を提供することにある。

【0008】

【課題を解決するための手段】本発明に係る請求項1記載の単語分類処理方法は、複数の単語を含むテキストデータに対して、互いに異なるすべての複数v個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数v個のクラスに割り当てるとステップと、上記複数v個のクラスの単語のうち出現頻度が高いv個未満の(c+1)個のクラスの単語を1つのウィンドウ内のクラスの単語として第1の記憶装置に記憶するステップと、上記第1の記憶装置に記憶された1つのウィンドウ内のクラスの単語に基づいて、互いに異なる第1のクラスの単語と第2のクラスの単語とが隣接して出現する確率を、上記第1のクラスの単語の出現確率と第2のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数c個のクラスに分類し、分類された複数c個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数c個のクラスとして第2の記憶装置に記憶するステップと、上記第2の記憶装置に記憶された中間層の複数c個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を

二分木の形式で1個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第3の記憶装置に記憶するステップと、上記第2の記憶装置に記憶された中間層の複数c個のクラスの各クラス毎に、上記中間層の複数c個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第4の記憶装置に記憶するステップと、上記第4の記憶装置に記憶された上記ツリー図の下側層を、上記第2の記憶装置に記憶された上記中間層の複数c個のクラスと連結する一方、上記第3の記憶装置に記憶された上記ツリー図の上側層を、上記第2の記憶装置に記憶された上記中間層の複数c個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第5の記憶装置に記憶するステップとを備えたことを特徴とする。

【0009】また、請求項2記載の単語分類処理方法は、請求項1記載の単語分類処理方法において、上記分類された複数c個のクラスを上記第2の記憶装置に記憶するステップは、上記第1の記憶装置に記憶された1つのウィンドウよりも外側のクラスが存在し、又は上記1つのウィンドウ内のクラスがc個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする。

【0010】本発明に係る請求項3記載の単語分類処理装置は、複数の単語を含むテキストデータに対して、互いに異なるすべての複数v個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数v個のクラスに割り当てる第1の制御手段と、上記複数v個のクラスの単語のうち出現頻度が高いv個未満の(c+1)個のクラスの単語を1つのウィンドウ内のクラスの単語として第1の記憶装置に記憶する第2の制御手段と、上記第1の記憶装置に記憶された1つのウィンドウ内のクラスの単語に基づいて、互いに異なる第1のクラスの単語と第2のクラスの単語とが隣接して出現する確率を、上記第1のクラスの単語の出現確率と第2のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数c個のクラスに分類し、分類された複数c個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数c個のクラスとして第2の記憶装置に記憶する第3の制御手段と、上記第2の記憶装置に記憶された中間層の複数c個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第3の記憶装置に記憶する第4の制御手段と、上記第2の記憶装置に記

憶された中間層の複数c個のクラスの各クラス毎に、上記中間層の複数c個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第4の記憶装置に記憶する第5の制御手段と、上記第4の記憶装置に記憶された上記ツリー図の下側層を、上記第2の記憶装置に記憶された上記中間層の複数c個のクラスと連結する一方、上記第3の記憶装置に記憶された上記ツリー図の上側層を、上記第2の記憶装置に記憶された上記中間層の複数c個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第5の記憶装置に記憶する第6の制御手段とを備えたことを特徴とする。

【0011】また、請求項4記載の単語分類処理装置は、請求項3記載の単語分類処理装置において、上記第3の制御手段は、上記第1の記憶装置に記憶された1つのウィンドウよりも外側のクラスが存在し、又は上記1つのウィンドウ内のクラスがc個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することを特徴とする。

【0012】本発明に係る請求項5記載の音声認識装置は、入力される発声音の音声信号に基づいて、請求項3又は4記載の単語分類処理装置によって複数の単語が複数のクラスに分類された単語分類結果を含む単語辞書と、所定の隠れマルコフモデルとを参照して上記発声音を音声認識する音声認識手段を備えたことを特徴とする。

【0013】

【発明の実施の形態】以下、図面を参照して本発明に係る実施形態について説明する。図1は、本発明に係る第1の実施形態の音声認識装置のブロック図である。この音声認識装置は、テキストデータメモリ10内のテキストデータ内の単語について出現頻度の比較的低い単語を、同一の単語に隣接する割合の多い単語を同一のクラスに割り当てるという基準で分類した後、単語分類結果を中間層、上側層、及び下側層の3つの階層に分類し、テキストデータ内のすべての単語を対象とするグローバルな(全体的な)コスト関数である所定の平均相互情報量を用いて、中間層、上側層、及び下側層の順序で階層別に単語の分類を実行して、単語辞書メモリ11内に単語辞書として格納する単語分類処理部20を備えたことを特徴とする。

【0014】＜単語分類処理方法＞まず、本発明に係る本実施形態の単語の分類(クラスターリング)方法について、第1の従来例の方法と対比させて説明する。本発明の方法は、従来技術文献に開示された第1の従来例の方

法を修正しかつ大幅に発展させて改善させた方法であって、第1の従来例の式と、本実施形態の式との相違について説明し、次いで、単語の分類処理方法について説明する。ここで、第1の従来例と、本実施形態とを比較するために、第1の従来例で用いた表記法と同一の表記法を用いることにする。

【0015】まず、相互情報量を用いたクラスタリングの方法について述べる。ここで、単語数Tのテキスト、語数Vの語彙、それに語彙の分割関数 π が存在すると仮定し、ここで、語彙の分割関数 π は語彙Vから語彙の中の単語クラスセットCへの分割写像（マッピング）を表わす写像関数である。第1の従来例においては、複数

$$\begin{aligned} I &= \sum_{C_1, C_2} \Pr(C_1, C_2) \log \{ \Pr(C_1 | C_2) / \Pr(C_1) \} \\ &= \sum_{C_1, C_2} \Pr(C_1, C_2) \log \{ \Pr(C_1, C_2) / (\Pr(C_1) \cdot \Pr(C_2)) \} \end{aligned}$$

【0019】ここで、 $\Pr(C_1)$ は第1のクラス C_1 の単語の出現確率であり、 $\Pr(C_2)$ は第2のクラス C_2 の単語の出現確率であり、 $\Pr(C_1 | C_2)$ は、第2のクラス C_2 の単語は出現した後に、第1のクラス C_1 の単語が出現する条件付き確率であり、 $\Pr(C_1, C_2)$ は第1のクラス C_1 の単語と第2のクラス C_2 の単語が隣接して出現する確率である。従って、上記数2で表されるAMIは、互いに異なる第1のクラス C_1 の単語と第2のクラス C_2 の単語とが隣接して出現する確率を、上記第1のクラス C_1 の単語の出現確率と第2のクラス C_2 の単語の出現確率との積で割った相対的な頻度の割合を表わす。

【0020】エントロピーHは写像関数 π に依存しない値であることから、AMIを最大にする写像関数は同時にテキストの尤度 $L(\pi)$ も最大にする。従って、AMIを単語のクラス構成における目的関数として使用することができる。

【0021】第1の従来例の相互情報量を用いたクラスタリング方法では、下側層から上側層へのボトムアップのマージ手順を用いている。初期の段階では、各単語をそれぞれ1つのクラスに割り当てる。次いで、すべてのクラスのペア（対）の中で最小のAMIの減少量を与える2つのクラスのペアを探索し、その2つのクラスのペアをマージし、マージ後のクラス数が予め決められた数cになるまで上記マージの処理を繰り返す。この第1の従来例の基本的な方法において、例えばコンピュータによって実行される演算時間のコンプレキシティー（又は演算時間のコスト）は、当該処理を以下に示すように直接的に実行したとき、 V^5 （語彙の語数Vの5乗）に比例するオーダーであり、これを $O(V^5)$ と表記する。ここで、演算時間のコンプレキシティーは、演算時間がどれぐらいかかるかを示す指標である。

【0022】＜ステップA1＞マージ処理の回数は合計で $(V-c)$ 回であり、このときの演算時間のコンプレ

キシティーは単語からなるテキストデータを生成するバイグラムのクラスモデルの尤度 $L(\pi)$ は次式によって得られる。

【0016】

【数1】 $L(\pi) = -H + I$

【0017】ここで、Hはモノグラムの単語分布のエントロピーであり、Iはテキストデータ内の隣接する2つのクラス C_1, C_2 に関する平均的な相互情報量（Average Mutual Information；以下、平均相互情報量とし、AMIと表記する。）であり、次式で計算することができる。

【0018】

【数2】

キシティーは語彙の語数Vに比例するオーダー $O(V)$ である。

＜ステップA2＞n回のマージ処理の後には、 $(V-n)$ 個のクラスが残り、次のマージ処理の段階では、組み合わせ数 ${}_{V-n}C_2$ （すなわち、 $V-n$ 個のクラスから2つのクラスをとるときの組み合わせ数）個のマージ処理のテスト又はトライアル（trial）（以下、トライアルといい、ここで、複数回のマージ処理を実行するが、実際にマージして単語分類結果に反映させるのは、このうちの1つであるので、本実施形態ではこのように呼ぶ。）を実行して探索する必要がある。そのうちの1つのみが後のマージ処理で有効化される。従って、このときの演算時間のコンプレキシティーは語彙の語数の2乗 V^2 に比例するオーダー $O(V^2)$ である。

＜ステップA3＞第n段階での1つのマージ処理のトライアルには、上記数2を用いてAMIを演算するための $(V-n)^2$ 個の項又はクラスに関する加算演算を含む。従って、このときの演算時間のコンプレキシティーは語彙の語数の2乗 V^2 に比例するオーダー $O(V^2)$ である。

【0023】従って、例えばコンピュータによって実行される全体の演算時間のコンプレキシティーは、語数の5乗 V^5 に比例するオーダー $O(V^5)$ となる。しかしながら、後述するように、冗長的な計算を除くことによって、演算時間のコンプレキシティーを、語彙の語数の3乗 V^3 に比例するオーダー $O(V^3)$ に減らすことも可能である。つまり、次のような本発明に係る方法によれば、上記ステップA3の部分を一定時間で実行することができる。

【0024】＜ステップB1＞上記数2は、前の段階のマージ処理で値が変更されたクラスのみについて計算する。従って、演算時間のコンプレキシティーは、第1の従来例におけるオーダー $O(V^2)$ から、オーダー $O(V)$ となる。

＜ステップB2＞前の段階のマージ処理におけるすべてのトライアルの結果を格納する。従って、演算時間のコンプレキシティーは、第1の従来例におけるオーダーO(V)から、語彙の語数Vに依存しない一定のオーダーO(1)となる。

【0025】例えば、V個のクラス数の語彙から始めて既に(V-k)回のマージ処理を実行して、k個のクラス $C_k(1)$, $C_k(2)$, ..., $C_k(k)$ が残っていると仮定する。この段階でのAMI I_k は次式で計算される。

【0026】

【数3】

$$I_k = \sum_{l, m} q_k(l, m)$$

【数4】 $q_k(l, m) = p_k(l, m) \log [p_k(l, m) / \{p_l(l) p_r(m)\}]$

【0027】ここで、 $p_k(l, m)$ は、クラス $C_k(1)$ における単語の次に、クラス $C_k(m)$ における単語が続く確率であり、次式のように表される。なお、本明細書及び図面において、表示を明確にするために、l(小文字のエル)としてlをも用い、 $l=1$ とする。

【0028】

【数5】

$$p_k(l, m) = P_r(C_k(l), C_k(m))$$

【数6】

$$\begin{aligned} p_l(l) &= \sum_m p_k(l, m) \\ p_r(m) &= \sum_l p_k(l, m) \end{aligned}$$

【0029】上記数3においては、 q_k は($k \times k$)クラスのバイグラム平面テーブルの全体にわたって加算さ

$$\begin{aligned} L_k(l, j) &= s_k(l) + s_k(j) - q_k(l, j) - q_k(j, l) \\ &\quad - \left(\sum_{l \neq 1, j} q_k(l, l+j) + \sum_{m \neq 1, j} q_k(l+j, m) + q_k(l+j, l+j) \right) \end{aligned}$$

ここで、

【数8】

$$s_k(l) = \sum_m q_k(l, l) + \sum_m q_k(l, m) - q_k(l, l)$$

【0032】すべてのクラスのペアのAMIの減少量 L_k を算出したら、当該AMIの減少量 L_k が最小となるようなペア、例えばクラス $C_k(i)$ とクラス $C_k(j)$ (但し $i < j$)とを選択し、次いで、これらのクラスのペアをマージさせたときの新しいクラスの名前を $C_{k-1}(i)$ と命名し、さらに、($k-1$)個のクラスの新たなセットによる次のマージ処理を続けて実行する。クラス $C_k(i)$ とクラス $C_k(j)$ を除くすべてのクラスについてマージ処理後に同じ方法で索引番号(インデックス)を付与する。すなわち、クラス $C_k(m)$ をクラス

れ、ここで、(l, m)のセルは $q_k(l, m)$ で表わす。いま、クラス $C_k(i)$ とクラス $C_k(j)$ とのマージ処理のトライアルを探索したとき、当該マージ処理によるAMIの減少量を、 $L_k(i, j) \equiv I_k - I_{k-1}(i, j)$ とし、ここで、 $I_k(i, j)$ は当該のマージ処理後のAMIである。

【0030】図3は、本発明に係る単語分類処理における加算領域及び加減算処理を示すクラスバイグラム平面テーブルの図である。ここで、図3及び、以下に示す図4と図5は、2つのクラスのバイグラムの平面を示す。図3に示すように、上記数3における加算領域P0は、図3の部分領域P1、P2及びP3の和から部分領域P4を減じた部分として表わすことができる。この4つの部分P1、P2、P3、P4のうち、部分領域P1の加算値は $C_k(i)$ と $C_k(j)$ とのマージ処理によって変化することはない。従って、AMIの減少量 $L_k(i, j)$ を算出する場合、加算領域P0を、2次元の領域(すなわち、正方形の領域)から1次元の領域(すなわち、複数のライン又は線)に減らすことが可能である。よって、上記ステップA3における演算時間のコンプレキシティーは、オーダーO(V^2)からオーダーO(V)に減少させることができる。クラス $C_k(i)$ とクラス $C_k(j)$ とのマージ処理によって生成されるクラスを表わす表記法として、 $C_k(i+j)$ を使用すると、AMIの減少量 $L_k(i, j)$ は次式によって与えられる。

【0031】

【数7】

$C_{k-1}(m)$ とし、ただし、 $m \neq i, j$ である。ここで、 $j \neq k$ であればクラス $C_k(k)$ をクラス $C_{k-1}(j)$ とし、 $j = k$ であればマージ処理後に $C_k(k)$ を削除する。

【0033】前の段階のマージ処理によるすべてのAMIの減少量 L_k を記憶装置に格納することによって、別の最適化処理を実行することができる。ここで、クラスのペア($C_k(i)$, $C_k(j)$)がマージ処理の対象として選択され、すなわち、すべてのペア(l, m)に対して、 $L_k(i, j) \leq L_k(l, m)$ であると仮定する。次のマージ処理の段階では、すべての(l, m)のペアに対して、AMIの減少量 $L_{k-1}(l, m)$ を計算する必要がある。ここで、上付き文字(i, j)は、クラスのペア($C_k(i)$, $C_k(j)$)が前のマージ処理の段階でマージされたことを意味している。ここ

で、AMIの減少量 $L_{k-1}^{(i,j)}(l, m)$ と $L_k(l, m)$ の違いに注意する必要がある。すなわち、 $L_{k-1}^{(i,j)}(l, m)$ はクラスiとクラスjとのマージ処理の後にクラスlとクラスmとをマージしたことによるAMIの減少量であり、 $L_k(l, m)$ はクラスiとクラスjとのマージ処理なしにクラスlとクラスmとをマージしたことによるAMI減少量である。従って、AMIの減少量 $L_{k-1}^{(i,j)}(l, m)$ と、AMIの減少量 $L_k(l, m)$ との差分は、クラスのペア $(C_k(i), C_k(j))$ のマージ処理によって影響を受ける項又はクラスのみから発生する。

【0034】上記の処理を、図4を参照して説明すると、AMIの減少量 $L_{k-1}^{(i,j)}(l, m)$ と、AMIの減少量 $L_k(l, m)$ に対するクラスバイグラム平面テーブルの加算領域は図4の(b)及び(a)のようになる。領域 $\{(x, y) \mid x \neq i, j, l, m, \text{及び} y \neq i, j, l, m\}$ の加算値は、クラスiとクラスjとのマージ処理によって、あるいは、クラスlとクラスmとのマージ処理によって変化することはないため、それらの領域については図示していない。さらに、詳細後述するように、 $\{L_{k-1}^{(i,j)}(l, m) - L_k(l, m)\}$ を計算するときに、図4の図中のほとんどの領域は互いに相殺されて数カ所のポイントの領域のみが残る。こうして、上記ステップA3における演算時間のコンプレキシティーを定数にまで減少することができる。

【0035】

【数9】 $L_k(l, m) = I_k - I_k(l, m)$

及び

$L_{k-1}^{(i,j)}(l, m) = I_{k-1}^{(i,j)} - I_{k-1}^{(i,j)}(l, m)$,

であるので、

$L_{k-1}^{(i,j)}(l, m) - L_k(l, m) = -(I_{k-1}^{(i,j)}(l, m) - I_k(l, m)) + (I_{k-1}^{(i,j)} - I_k)$

【0036】AMIの減少量 $I_{k-1}^{(i,j)}(l, m)$ と、AMIの減少量 I_k の加算領域のうちの幾つかの部分は、AMIの減少量 $I_{k-1}^{(i,j)}$ の一部、あるいはAMIの減少量 $I_k(l, m)$ の一部とともに相殺される。ここで、 $I_{k-1}^{(i,j)}(l, m)$ 、 $I_{k-1}^{(i,j)}$ 、 $I_k(l, m)$ 、 I_k はそれぞれ、相殺可能な共通のクラスをすべて相殺した後のAMIの減少量 $I_{k-1}^{(i,j)}(l, m)$ 、 $I_k(l, m)$ 、 $I_{k-1}^{(i,j)}$ 、 I_k であることを表わす。このとき、次のような関係式が与えられる。

【0037】

【数10】 $L_{k-1}^{(i,j)}(l, m) - L_k(l, m) = -(I_{k-1}^{(i,j)}(l, m) - I_{k-1}^{(i,j)}) + (I_{k-1}^{(i,j)} - I_k)$

ここで、

【数11】 $I_{k-1}^{(i,j)}(l, m) = q_{k-1}(l+m, i) + q_{k-1}(i, l+m)$

【数12】 $I_{k-1}^{(i,j)}(l, m) = q_k(l+m, i) + q_k(i, l+m)$

$+m) + q_k(l+m, j) + q_k(j, l+m)$

【数13】 $I_{k-1}^{(i,j)}(l, m) = q_{k-1}(i, l) + q_{k-1}(i, m) + q_{k-1}(l, i) + q_{k-1}(m, i)$

【数14】 $I_k(l, m) = q_k(i, l) + q_k(i, m) + q_k(j, l) + q_k(j, m) + q_k(l, i) + q_k(l, j) + q_k(m, i) + q_k(m, j)$

【0038】上記数10における I_h の加算領域を図5に示す。第1の従来例においては、上記数10の右辺第2項を無視して第1項のみを使用して、AMIの減少量 $L_{k-1}^{(i,j)}(l, m) - L_k(l, m)$ を計算しているようである。なお、上記数10の第1項に対応する従来技術文献における方程式(17)の第3式において、符号は正負逆である。しかしながら、上記数10の第2項は、その第1項と同じ重み係数を有するので、本発明者は、本発明の当該モデルを完全なものとするために、上記数10を用いる。

【0039】演算時間のコンプレキシティーのオーダー $O(V^3)$ を有する方法を使用する場合でも、語彙数が 10^4 又はそれ以上のオーダーのように大きいときには、実際に計算することができない。何れにしても、上記ステップA1において、オーダー $O(V)$ の演算時間が必要であるため、修正できるのは上記ステップA2しかないと考えられる。上記ステップA2においては、可能なクラスペアのマージのすべてについて検討することもできるが、実際には探索するクラスペアの範囲を限定することは可能である。このことに関しては、第1の従来例においては、以下のような方法を提案しており、本発明の方法もこれを採用している。まず、互いに重複しない単語数 V を含むテキストデータ内の単語に基づいて、 V 個の単体のクラスを作り、これを頻度の高い順に配列して、「マージ領域」(本実施形態では、ウィンドウという。従って、本明細書においては、マージ領域とウィンドウとは同義語である。)を、クラス順位の最初の $c+1$ 個のクラスの単語とする。従って、まずは、 $(c+1)$ 個の頻度の高い単語がマージ領域となる。次いで下記の処理を行う。

【0040】<ステップD1>マージ領域内のすべてのペアの中でも、AMIの減少量を最小にするようなクラスのペアをマージする。

<ステップD2> $(c+2)$ 番目の位置にあるクラスをマージ領域又はウィンドウの中に挿入し、 $(c+2)$ 番目の位置のクラスよりも後ろの各クラスをその左側方向に1つだけ移動させる。

<ステップD3>残りのクラスが所定の c 個になるまで上記ステップD1とD2の処理を繰り返す。

【0041】当該第1の従来例の処理のアルゴリズムにおいては、上記ステップA2の演算時間のコンプレキシティーは、最終クラス数 c の2乗である c^2 に比例するオーダー $O(c^2)$ となり、全体の演算時間は、 c^2V に比例するオーダー (c^2V) に減少する。

【0042】次いで、単語のクラスタリング構造を得るための方法について述べる。単語のクラスタリング構造を表わすツリーでの表現を得るための最も簡単な方法は、マージ処理における副産物としてデンドログラム

(dendrogram; ツリーの系統図又はツリー図ともいう。) 系統樹を構築すること、即ち具体的には、マージの順序の記録(又は履歴)を取ってその記録に基づいて二分木を作ることである。図6に、5単語から成る語彙を使用した簡単な例を示す。図6におけるマージ履歴(又はマージのヒストリともいう。)は、次の表1に示す通りである。なお、表1の第1行目は、「クラスAとクラスBとをマージして、マージ後の新しいクラスをAと名づけた。」ということを意味する。

【0043】

【表1】

マージ履歴

Merge (A, B→A)
Merge (C, D→C)
Merge (C, E→C)
Merge (A, C→A)

【0044】しかしながら、この方法を、上記第1の従来例の方法のO(C²V)アルゴリズムに直接的に適用した場合、各クラスのバランスは極端に悪くなり、図7に示されているようなほぼ左側方向の分岐のみのツリー構造となる。この理由は、AMI量に関して言えば、マージ領域にある複数のクラスをある一定の大きさを有するように成長させた後に、比較的大きなサイズを有するより高い頻度を有するクラスをマージするよりは、より低い頻度を有する単集合のクラスをマージした方が、大幅にコストが安くなるからである。

【0045】本発明者が採用した本発明に係る新しい方法は以下の通りである。

<ステップE1>MIクラスタリング: マージ領域の制約条件を有する相互情報クラスタリングアルゴリズムを使用してc個のクラスを作成する。当該c個のクラスは、図19に示すように、最後のツリー図であるデンドログラムの中間層100を構成する。

<ステップE2>外部クラスタリング: テキストデータ中のすべての単語をクラス・トークン(class token)と置換し(実際には、テキストデータ中の全体の文章の代わりにパイグラム・テーブルについてのみ処理を行う。)、マージ領域の制約条件なしにすべてのクラスがマージ処理によって単一のクラスになるまで二分木の形式でマージ処理(バイナリーマージ処理)を実行する。当該処理によって、デンドログラムD_{root}を作成する。このデンドログラムD_{root}は、例えば図19に示すように、最終のツリー構造の上側層101を構成する。

<ステップE3>内部クラスタリング: {C¹, C², ..., Cⁱ, ..., C^c}を上記ステップE1で得られた中間層100のクラスの集合(クラスセット)とする。そして、それぞれのi(1 ≤ i ≤ c; iは自然数である。)について以下の処理を行う。

【0046】<ステップE3-1>クラスC^jのものを除いて、テキスト中のすべての単語をそのクラス・トークンと置き換える。新しい語彙V' = V₁ ∪ V₂を決定する。ここで、V₁ = {C^jにおけるすべての単語}、V₂ = {C¹, C², ..., Cⁱ⁻¹, Cⁱ⁺¹, C^c}であり、C^jはj番目のクラスのトークンである(1 ≤ j ≤ c)。語彙V'の各要素を各々のクラスに割り当て、語彙V₁の要素のみを含むクラスに限ってマージ処理が可能となるという制約条件付きマージ処理によって二分木の形式でマージ処理を実行する。当該処理は、最初の|V₁|個のクラスにおける語彙V₁の要素(すなわち、単語)を含む語彙V'の要素を頻度の順序で順序づけし、次いで、最初に幅|V₁|を有しかつ各マージ処理によって1つずつ減少する幅を有するマージ領域内でマージ処理を実行することによって実行することができる。ここで、|V₁|は語彙V₁の単語の個数である。

<ステップE3-2>語彙V₁におけるすべての要素が単一のクラスに入るまでマージ処理を繰り返す。各クラス毎に、マージ処理によって、図19に示すように、下側層102のデンドログラムD_{sub}を作成する。このデンドログラムは、葉のノード(leaf node)がクラス内の各単語を表している各クラスのサブツリーを構成する。

【0047】<ステップE4>上側層101のデンドログラムD_{root}の各葉のノードを、対応する下側層102のデンドログラムD_{sub}と置き換えることによってデンドログラムを合成し、これによって、全体のデンドログラムを得ることができる。

【0048】本発明に係る単語クラスタリングの方法は、意味又は統語的特徴が似通った単語が近接した位置に配置された点で、バランスが取れた二分木の形式を有するツリー構造を生成することができる。図8は、本発明の方法を用いて、ウォール・ストリート・ジャーナル(以下、WSJという。)のコーパスの中で最も使用頻度の高い上位70,000語に関して組み立てた500クラスの内の1クラスに対する下側層のデンドログラムD_{sub}の一例を示したものである。最後に、根のノード(ルートノード(root node))から葉のノード(リーフノード(leaf node))に至るパスの追跡し、左側方向の分岐又は右側方向の分岐をそれぞれ表わす0又は1の1ビットを各分岐に割り当てることによって、語彙の中の各単語に対して、ビットストリング(単語ビット)を割り当てることができる。

【0049】図10は、図1の単語分類処理部20の構成を示すブロック図である。図10を参照して、単語分

類処理部 20 の構成及び動作について説明する。図 10 において、単語分類処理部 20 は、CPU 50 を備えたコントローラであって、CPU 50 と、CPU 50 によって実行される単語分類処理のプログラム及び当該プログラムを実行するために必要なデータを格納するための ROM 51 と、上記単語分類処理を実行するときに必要なワークエリアであるワーク RAM 52 と、上記単語分類処理を実行するときに必要な複数のメモリエリアを有する RAM 53 と、2 つのメモリインターフェース 54、55 とを備え、これらの各回路 50 乃至 55 はバス 56 を介して互いに接続される。ここで、メモリインターフェース 54 は、テキストデータメモリ 10 と CPU 50 との間に設けられ、テキストデータメモリ 10 と CPU 50 との間の信号変換などのインターフェース処理を実行するためのインターフェース回路である一方、メモリインターフェース 55 は、単語辞書メモリ 11 と CPU 50 との間に設けられ、単語辞書メモリ 11 と CPU 50 との間の信号変換などのインターフェース処理を実行するためのインターフェース回路である。RAM 53 は、次のように区分された複数のメモリ部を備える。

【0050】(a) 初期化クラス単語メモリ 61 : 後述する初期化処理によって得られた v 個の単語及びそのクラスを格納する；

(b) AMI メモリ 62 : 後述する中間層クラスタリング処理、上側層クラスタリング処理及び下側層クラスタリング処理において 1 つのウィンドウ内のクラスの単語の中ですべての組み合わせの仮ベアを作り、各仮ベアをマージしたときの平均相互情報量を数 2 を用いて計算した結果を格納する；

(c) 中間層メモリ 63 : 後述する中間層クラスタリング処理によって得られた c 個の中間層のクラスの単語を格納する；

(d) 上側層ヒストリメモリ 64 : 後述する上側層クラスタリング処理における各マージ処理の履歴（又はヒストリ）を格納する；

(e) 上側層ツリーメモリ 65 : 上記上側層クラスタリング処理によって得られたツリー図であるデンドログラム D_{root} を格納する；

(f) 下側層ヒストリメモリ 66 : 上記下側層クラスタリング処理によって得られた、中間層 100 の各クラスに対して 1 つのツリー図である複数の c 個のデンドログラム D_{sub} を格納する；

(g) 下側層ツリーメモリ 67 : 上記下側層クラスタリング処理によって得られたツリー図であるデンドログラム D_{sub} を格納する；

(h) ツリーメモリ 67 : 上側層 101 の 1 つのデンドログラムと下側層 102 の複数の c 個のデンドログラムとを、中間層 100 を介して連結することにより得られた全体のツリー図であるデンドログラムを格納する。

【0051】図 11 は、図 1 の単語分類処理部 20 によ

って実行されるメインルーチンの単語分類処理を示すフローチャートである。図 11 に示すように、まず、ステップ S1 においてテキストデータに基づいて出現頻度の高い単語から順に並べる処理を実行する初期化処理を実行し、次いで、ステップ S2 において中間層 100 のクラスの単語を求める中間層クラスタリング処理を実行し、さらに、ステップ S3 において上側層 101 のツリー図を求める上側層クラスタリング処理を実行し、そして、ステップ S4 において下側層 102 のツリー図を求める下側層クラスタリング処理を実行し、最後に、ステップ S5 において上側層 101 の 1 つのツリー図と下側層 102 の複数の c 個のツリー図とを、中間層 100 を介して連結することにより得られた全体のツリー図であるデンドログラムを求めて、その結果を単語辞書として単語辞書メモリ 11 に格納するデータ出力処理を実行する。これによって、単語分類処理が終了する。なお、これらのツリー図においては、各単語がそれぞれ 1 つのクラスに分類されかつクラス間の連結関係が示される。

【0052】図 12 は、図 11 のサブルーチンの初期化処理（S1）を示すフローチャートである。図 12 に示すように、ステップ S11 において、テキストデータメモリ 10 内のテキストデータに基づいて、単語の重複を省いたすべての複数の v 個の単語の出現頻度を調べて、出現頻度の高い単語から順に並べて、これを複数の v 個のクラスに割り当てて、複数の v 個のクラスの単語を初期化クラス単語メモリ 61 に記憶して、元のメインルーチンに戻る。ここで、 v は 2 以上の自然数である。

【0053】図 13 は、図 11 のサブルーチンの中間層クラスタリング処理（S2）を示すフローチャートである。図 13 に示すように、まず、ステップ S21 において、初期化クラス単語メモリ 61 から複数の v 個のクラスの単語を読み出した後、複数の v 個のクラスの単語のうちの出現頻度の高いクラスの単語から v 個未満の $(c+1)$ 個のクラスの単語を 1 つのウィンドウ（又はマージ領域）内のクラスの単語として、図 17 に示すように、ワーク RAM 52 に記憶する。ここで、 $1 < c+1 < v$ である。次いで、ステップ S22 において、ワーク RAM 52 に記憶された 1 つのウィンドウ内のクラスの単語の中で、すべての 2 個ずつの組み合わせの仮ベアを作り、各仮ベアをそれぞれマージしたときの平均相互情報量を数 2 を用いて計算して、各仮ベアとそれに対応する計算された平均相互情報量とを次の表 2 の形式で AMI メモリ 62 に記憶する。

【0054】

【表 2】

仮ベア	平均相互情報量
(C^1, C^1)	0.867678
(C^2, C^3)	0.234689

(C³, C⁴) 0. 1 2 5 6 8 6

(C⁵, C⁶) 0. 6 7 5 6 4 2

【0055】次いで、ステップS23において、図18に示すように、AMIメモリ62に記憶された各仮ベアの平均相互情報量のうち、最大となる仮ベアを見つけて当該仮ベアをマージすることにより、1つのクラスが減少し、マージ後の1つのウィンドウ内のクラスの単語を更新して、更新後のクラスの単語をワークRAM52に記憶する。そして、ステップS24において、ウィンドウ外のクラスはなくなりかつウィンドウ内のクラスの数 c 個となったか否かが判断され、その判断がNOであるとき、ステップS25において、図18に示すように、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語をウィンドウ内に挿入し、挿入後の1つのウィンドウ内のクラスの単語を更新して、更新後のクラスの単語をワークRAM52に記憶した後、ステップS22に戻って、ステップS22以降の処理を繰り返す。

【0056】一方、ステップS24においてYESであるときは、ステップS26において、ワークRAM52に記憶された、ウィンドウ内の c 個のクラス及びそれに属する単語を中間層100として中間層メモリ63に記憶する。これによって、中間層クラスタリング処理が終了し、メインルーチンに戻る。

【0057】図14は、図11のサブルーチンの上側層クラスタリング処理(S3)を示すフローチャートであり、図19に示すように、中間層100から矢印201の方向でツリー図を求める処理である。図14に示すように、まず、ステップS31において、中間層メモリ63内の c 個のクラスの単語を読み出した後、当該 c 個のクラスの単語を1つのウィンドウ内のクラス単語として、ワークRAM52に記憶する。次いで、ステップS32において、ステップS22と同様に、ワークRAM52に記憶された1つのウィンドウ内のクラスの単語の中で、すべての2個ずつの組み合わせの仮ベアを作り、各仮ベアをそれぞれマージしたときの平均相互情報量を数2を用いて計算して、各仮ベアとそれに対応する計算された平均相互情報量とを前述の表2の形式でAMIメモリ62に記憶する。

【0058】次いで、ステップS33において、ステップS23と同様に、AMIメモリ62に記憶された各仮ベアの平均相互情報量のうち、最大となる仮ベアを見つけて当該仮ベアをマージすることにより、1つのクラスが減少し、マージ後の1つのウィンドウ内のクラスの単語を更新して、更新後のクラスの単語をワークRAM52に記憶する。また、例えば表1の形式を有し、どのクラスとどのクラスとがマージされて新しく名づけられたクラスとなったかを表わす当該マージ処理の履歴を上側

層ヒストリメモリ64に記憶する。そして、ステップS34において、ウィンドウ内のクラスの数 c 個となったか否かが判断され、その判断がNOであるとき、ステップS32に戻って、ステップS32以降の処理を繰り返す。

【0059】一方、ステップS34においてYESであるときは、ステップS35において、上側層ヒストリメモリ64内の上側層の履歴又はヒストリに基づいて、例えば図6に示すように、上側層のツリー図又はデンドログラム D_{root} を作成して上側層ツリーメモリ65に記憶する。これによって、上側層クラスタリング処理が終了し、メインルーチンに戻る。

【0060】図15は、図11のサブルーチンの下側層クラスタリング処理(S4)を示すフローチャートであり、図15に示すように、下側層102の底辺に位置する単語から、中間層100の各クラス C_i 毎に、矢印202の方向でツリー図を求める処理である。ある。図15に示すように、まず、ステップS41において、中間層メモリ63内の c 個のクラスの単語を読み出した後、当該 c 個のクラスから1つのクラスを選択する。そして、ステップS42において、選択されたクラス内の v_i 個の単語を1つのウィンドウ内のクラス単語として、ワークRAM52に記憶する。次いで、ステップS43において、ステップS22及びS32と同様に、ワークRAM52に記憶された1つのウィンドウ内のクラスの単語の中で、すべての2個ずつの組み合わせの仮ベアを作り、各仮ベアをそれぞれマージしたときの平均相互情報量を数2を用いて計算して、各仮ベアとそれに対応する計算された平均相互情報量とを前述の表2の形式でAMIメモリ62に記憶する。

【0061】次いで、ステップS44において、ステップS23及びS33と同様に、AMIメモリ62に記憶された各仮ベアの平均相互情報量のうち、最大となる仮ベアを見つけて当該仮ベアをマージすることにより、1つのクラスが減少し、マージ後の1つのウィンドウ内のクラスの単語を更新して、更新後のクラスの単語をワークRAM52に記憶する。また、例えば表1の形式を有し、どのクラスとどのクラスとがマージされて新しく名づけられたクラスとなったかを表わす当該マージ処理の履歴を下側層ヒストリメモリ66に記憶する。そして、ステップS45において、ウィンドウ内のクラスの数 c 個となったか否かが判断され、その判断がNOであるとき、ステップS43に戻って、ステップS43以降の処理を繰り返す。ここで、ステップS43及びS44の処理は、中間層100の各クラス毎に実行される。

【0062】一方、ステップS45においてYESであるときは、ステップS46においてすべての中間層100のクラスについて処理したか否かが判断され、当該判断がNOであるとき、未処理のクラスが残っているの、ステップS47において残っている中間層100の

別の未処理のクラスを選択した後、ステップS42に進む。一方、ステップS46においてYESであるときは、ステップS48において、下側層ヒストリメモリ66内の下側層の履歴又はヒストリに基づいて、例えば図6に示すように、下側層のツリー図又はデンドログラム D_{sub} を作成して下側層ツリーメモリ67に記憶する。これによって、下側層クラスタリング処理が終了し、メインルーチンに戻る。

【0063】図16は、図11のサブルーチンのデータ出力処理(S5)を示すフローチャートである。図16に示すように、まず、ステップS51において、図19に示すように、上側層ツリーメモリ65内の上側層のツリー図と、下側層ツリーメモリ67内の下側層のツリー図とに基づいて、これら2つのツリー図を中間層100の各クラスCⁱを介して連結し、すなわち、上側層ツリーメモリ65内の上側層のツリー図を中間層100の各クラスCⁱに連結する一方、下側層ツリーメモリ67内の下側層ツリー図をその頂点にあるクラスを中間層100の各クラスCⁱに連結する。これによって、当該テキストデータに基づく全体のツリー図を作成して、ツリー図の情報をツリーメモリ68に記憶する。当該ツリーメモリ68には、図6及び図8に示すように、各クラスの単語間の連結関係が単語辞書として記憶される。そして、ステップS52において、ツリーメモリ68内のツリー図の情報を単語分類結果(又は単語クラスタリング結果)として単語辞書メモリ11に出力して記憶する。

【0064】<第1の実施形態>図1は、本発明に係る第1の実施形態である音声認識装置のブロック図である。図1において、テキストデータメモリ10内に格納された、例えば英語又は日本語の複数の単語を含むテキストデータは、単語分類処理部20によって上述の単語分類処理が実行されることにより、複数のクラスに分類されかつクラスの連結関係が記述された単語辞書として、単語辞書メモリ11内に格納される。

【0065】一方、マイクロホン1に入力された複数の単語からなる発声音声は、マイクロホン1によって音声信号に変換された後、A/D変換器2によってデジタル音声信号にA/D変換される。デジタル音声信号は特徴抽出部3に入力され、特徴抽出部3は、入力されたデジタル音声信号に対して例えばLPC分析してケプストラム係数や対数パワーなどの特徴パラメータを抽出して、バッファメモリ4を介して音声認識部5に出力する。音声認識部5は、単語辞書メモリ11に格納された単語辞書を参照しかつ、例えば音素隠れマルコフモデル(以下、音素HMMという。)である言語モデルメモリ12に格納された言語モデルを参照して、単語毎に音声認識を実行して、音声認識結果を出力する。なお、ここで、単語辞書メモリ11内の単語辞書は、例えば、

(a) 010010010, position;

(b) 010010011, location;

(c) 110010100, for;

のように各単語とその単語の属するクラスを表現するビット列などの情報を含む。

【0066】<第2の実施形態>図2は、本発明に係る第2の実施形態である形態素及び構文解析装置のブロック図である。図2において、テキストデータメモリ31, 32にそれぞれ格納された、複数の単語からなる2つのテキストデータはそれぞれ、単語分類処理部20によって上述の単語の分類処理が実行されることにより、複数のクラスに分類されかつクラスの連結関係が記述された単語辞書として、それぞれ単語辞書メモリ41, 42内に格納される。

【0067】日本語又は英語などの所定の言語の文字列からなり複数の単語からなる自然言語文が形態素解析部21に入力され、形態素解析部21は、入力された自然言語文の各単語の出現形に対して、単語辞書メモリ41に格納された単語辞書を参照して上記自然言語文を複数の単語に分割するとともに、上記各出現形に対して品詞、活用形、標準表現形、及び類語コードなどの情報を付与し、これらの解析結果を構文解析部22に出力する。次いで、構文解析部22は、単語辞書メモリ42に格納された単語辞書を参照して、所定の構文解析を実行して単語列に対して構文木情報を付加して解析結果として出力する。

【0068】以上説明したように、図19に示すように、下側層、中間層、及び上側層と階層化して、複数の単語を二分木の形式で複数のクラスに分類したので、単語分類処理によりバランスのとれた階層構造を有する単語分類結果を得ることができる。また、AMIの計算においては、下側層、中間層、及び上側層ともに、すべてのクラスの単語を対象としてAMIを計算しているので、計算されたAMIは局所的なAMIではなく、全体の単語の情報を含んだグローバルAMIに基づいて、クラスタリング処理を実行している。従って、全体的に最適化された単語分類結果を得ることができる。これにより、テキストデータから単語の分類体系を自動的に獲得するときに、より精密で正確な分類体系を得ることができる。さらに、上記単語分類部20により得られた単語辞書に基づいて音声認識することにより、従来例に比較して高い認識率で音声認識することができる。

【0069】以上の実施形態において、音声認識部5と、単語分類処理部20と、形態素解析部21と、構文解析部22とは例えばデジタル計算機によって構成される。以上の実施形態の単語分類処理は、図11に示すように、中間層クラスタリング処理、上側層クラスタリング処理、下側層クラスタリング処理の順序で実行しているが、本発明はこれに限らず、中間層クラスタリング処理、下側層クラスタリング処理、上側層クラスタリング処理の順序で実行してもよい。以上の実施形態において、図11の初期化処理の前に、単語のn-グラムを利

用して、同一の単語に隣接する割合の多い単語を同一のクラスに割り当てるという基準で複数の単語を複数のクラスに分類する処理を実行してもよい。

【0070】

【実施例】

＜実験（シミュレーション）＞6年分のWSJのコーバスの平易なテキストデータを使用し、クラスと単語ビットを作成した。テキストのサイズは500万語、1000万語、2000万語、及び5000万語（それぞれ、5MW、10MW、20MW、50MW；ここで、Wはワードである。）である。語彙はコーバス全体で最も頻繁に使用されている上位7万語とした。最終クラス数 c は500に設定した。獲得したクラスと単語ビットを、それぞれ次の2つの尺度SS1とSS2を使用して評価する。

（a）尺度SS1は、WSJのコーバス、及び本出願人が所有するコーバスである一般的な英語のツリーバンクに基づいた、クラスを基本としたトライグラムモデルのパーブレキシティーを計算するパーブレキシティー法である。

（b）尺度SS2は、本出願人が所有する決定木を用いる部分音声のラベル付け（tagging又はlabeling）のラベル付け装置（tagger又はlabeler）における誤り率である。

【0071】＜パーブレキシティー法＞単語をその所属クラスに写像するクラス関数 G を使用すると単語トライグラムの確率は、次式のように書き直すことができる。

【0072】

$$[数15] P(w_i | w_{i-2}w_{i-1}) = P_c(G(w_i)) | G(w_{i-2})G(w_{i-1})P_m(w_i | G(w_i))$$

【0073】ここで、 P_c は2次のマルコフ連鎖確率であり、 P_m は単語メンバーシップ確率である。 P_c 及び P_m のスムージングは、それぞれカット（Katz）のバックオフ、及びグッドチューリング公式を使用して行う。トレーニング用テキストのサイズは1.9MWで、テストテキストは150KWであり、両者ともWSJの

コーバスを典拠としている。語彙サイズは77KWである。図9は、テストテキストのパーブレキシティーとクラスタリングのテキストサイズとの関係を示している。クラスタリングのテキストサイズにおけるゼロ点は、単語トライグラムモデルのパーブレキシティーを表している。クラスタリングのテキストサイズが増加するに従って、パーブレキシティーは単調に減少する。これはクラスタリング処理の改善を示している。50MWでは、パーブレキシティーは単語トライグラムモデルの場合より18%低くなっている。この結果は、クラス・トライグラムのパーブレキシティーが単語トライグラムモデルの場合より僅かに高いとした第1の従来例の結果とは好対照である。

【0074】＜決定木を用いた音声部分のラベル付け＞本出願人が所有する決定木を用いる部分音声のラベル付け（tagging）のラベル付け装置（tagger）は、スパッター（SPATTER、例えば、従来技術文献2「D. Magerman, "Natural Language Parsing as Statistical Recognition", Doctoral Dissertation, Stanford University, Stanford, California, 1994年」参照。）をベースとした、本出願人が所有する決定木パーザーの統合モジュールである。上記ラベル付け装置は、ユニバーシティ・オブ・ペンシルバニアのトリートバンクプロジェクトのそれよりも、1桁だけ大きい441個の統語的ラベル（syntactic tags）を採用している。学習用テキスト、テスト用テキスト、及び実行用テキストはすべて、単語とラベルとの対のすべてのシーケンスを含む。学習段階では、イベント（event）は、特徴値の集合又は、質問とそれに対する回答との対の集合である。1つの特徴は、処理すべき現在の単語word(0)が現れる文脈における任意の属性であり、これは便宜上、質問の形式で表される。ラベル付けは左から右へと行う。表3は、処理すべき現在の単語“like”を用いたイベントの一例を示している。

【0075】

【表3】

Event-128:

```
{
  <word(0), "like"><word(-1), "flies"><word(-2), "time">
  <word(1), "an"><word(2), "arrow"><tag(-1), "Verb-3rd-Sg-type3">
  >
  <tag(-2), "Noun-Sg-type14">
  .....
  <Inclass?(word(0), Class295), "yes"><WordBits(Word(-1), 29), "1">
  .....
  <Tag, "Prep-type5">
}
```

(Basic Questions)

(WordBits Questions)

【0076】このイベントの最後のペアは、回答、即ち

当該現在の単語の正しいラベルを示す特別な項目であ

る。最初の2行は当該現在の単語の回りの単語の識別に関する質問と、先行する単語のためのラベルを表している。これらの質問は、基本質問と呼ばれている。第2の質問形式(単語ビット質問)は、「この現在の単語はクラス295にありますか?」或いは「先行する単語の単語ビット中の第29ビットは何ですか?」と言ったクラスタ及び単語ビットに関するものである。

【0077】イベントの集合から決定木を作成する。決定木の根のノードは、それぞれ対応する単語に対して正しいラベルを含んでいるすべてのイベントからなるセットを表している。根のノード用のラベルの確率分布は当該集合におけるラベルの相対的な頻度を計算することによって得ることができる。当該セットの中の各イベントにおける特徴値を問い合わせることで、そのセットはN個のサブセットに分割することができる(Nは特徴に関する可能値である)。次いで、この特徴値を条件として、各サブセットに対するラベルの条件付き確率分布を計算することが可能である。セットの分割によって生じるエントロピーの減少を各特徴毎に計算した後、エントロピーの減少量を最大にする特徴を選択する。この方法を反復し、セットを各サブセットに分割することによって、葉のノードがタグの条件付き確率分布を含むような

決定木を構築することができる。次いで、獲得した確率分布を実行用データを使用してスムージングする。スムージング処理の詳細については上記従来技術文献2を参照せよ。テスト段階では、システムはテストテキスト内の各単語に対する条件付き確率分布を調査し、ビームサーチを使用して最も可能性のあるラベル付けシーケンスを選択する。

【0078】本発明者がラベル付け実験に使用したのは、WSJのテキスト、及び本出願人が所有するコーパス(以下、ATRコーパスという。)である。WSJのテキストは、本出願人の統語ラベルセットを使用して手動でラベル付けをし直した。上記ATRコーパスは、文語体の米語の包括的な見本であり、その語法のスタイル及び設定は非常に幅広く、多くの異なる領域から作り上げられている。ATRコーパスはまだ開発過程にあるため、この実験用として手元にあるテキストの大きさは、ラベルセットが大型である割にはかなり小型である。表4は今回の実験に使用したテキストのサイズを示している。

【0079】

【表4】

テキストサイズ(単語数)	学習用	テスト用	実行用
WSJのテキスト	75,139	5,831	6,534
ATRコーパス	76,132	23,163	6,680

【0080】図20は、多様なクラスタリングのテキストサイズに対するラベル付けの誤り率を表している。本実験では、2種類の質問形式の中から基本質問及び単語ビット質問を使用している。ラベル付け装置への単語ビット情報の導入の効果を見るため異なる実験を行ったが、その実験では無作為に生成されたビットストリングを各単語に割り当て(特徴的なビットストリングが各単語に割り当てられているが、ラベル付け装置もビットストリングを処理中の各単語の認識番号として使用している。この制御実験においては、ビットストリングの割り当ては無作為に行なわれるが、2つの単語が同じ単語ビットを持つことはない。無作為の単語ビットは、ラベル付け装置に対して単語の認識以外には何のクラス情報も与えない。)、基本質問と単語ビット質問を使用した。結果はクラスタリングのテキストサイズのゼロの値において表されている。WSJのテキスト及びATRコーパスの何れも、ラベル付けの誤り率は、5MWのテキストから抽出された単語ビット情報を使用することによって30%以上低下し、また、クラスタリングのテキストサイズが増加するとさらに誤り率が減少した。50MWでは、誤り率は43%も低下した。これもまた、クラスタの品質向上はクラスタリングのテキストサイズの増大に

よるものであることを示している。全体的にみて、高い誤り率は非常に大きなラベルセットと、小さな学習用セットに起因している。この結果の注目に値する点は、ATRコーパスのテキストとWSJのテキストは互いに領域が非常に異なったものであるにも関わらず、WSJのテキストから構成された単語ビットの導入が、WSJのテキストに対してと同じくらいATRコーパスのテキストのラベル付けにも効果的であったことである。この点から、獲得した階層的クラスタは領域を越えて移動可能であると考えられる。

【0081】以上説明したように、本発明者は、複数の単語の階層的クラスタリング分割に関するアルゴリズムを提案し、5MWから50MWまでの大型テキストデータを使用したクラスタ分割の実験を行った。獲得したクラスタの高品質性は、2種類の評価方法によって確認されている。クラスを基にしたトライグラムモデルのパーレキシティーは、単語をベースとしたトライグラムモデルの場合よりも18%低くなっている。本出願人が所有する決定木の音声部分のラベル付け装置に単語ビットを導入することにより、ラベル付けの誤りの割合は43%も減少する。WSJのテキストから得る階層的クラスタリング分割処理はまた、WSJのテキストとは全く異

なる範囲にあるATRコーパスのテキストのラベル付けにも有効であることが判った。

【0082】

【発明の効果】以上詳述したように本発明に係る請求項1記載の単語分類処理方法によれば、複数の単語を含むテキストデータに対して、互いに異なるすべての複数 v 個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数 v 個のクラスに割り当てるステップと、上記複数 v 個のクラスの単語のうち出現頻度が高い v 個未満の $(c+1)$ 個のクラスの単語を1つのウィンドウ内のクラスの単語として第1の記憶装置に記憶するステップと、上記第1の記憶装置に記憶された1つのウィンドウ内のクラスの単語に基づいて、互いに異なる第1のクラスの単語と第2のクラスの単語とが隣接して出現する確率を、上記第1のクラスの単語の出現確率と第2のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数 c 個のクラスに分類し、分類された複数 c 個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数 c 個のクラスとして第2の記憶装置に記憶するステップと、上記第2の記憶装置に記憶された中間層の複数 c 個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第3の記憶装置に記憶するステップと、上記第2の記憶装置に記憶された中間層の複数 c 個のクラスの各クラス毎に、上記中間層の複数 c 個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第4の記憶装置に記憶するステップと、上記第4の記憶装置に記憶された上記ツリー図の下側層を、上記第2の記憶装置に記憶された上記中間層の複数 c 個のクラスと連結する一方、上記第3の記憶装置に記憶された上記ツリー図の上側層を、上記第2の記憶装置に記憶された上記中間層の複数 c 個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として第5の記憶装置に記憶するステップとを備える。従って、下側層、中間層、及び上側層と階層化して、複数の単語を二分木の形式で複数のクラスに分類したので、単語分類処理によりバランスのとれた階層構造を有する単語分類結果を得ることができる。また、平均相互情報量の計算においては、下側層、中間層、及び上側層ともに、すべてのクラスの単語を対象として平均相互情報量を計算しているので、計算された平均相互情報量は局所的な平均相互情報量ではなく、全体の単語の情報を含んだグローバルな平均相互情報量に基づいて、クラスターリング処理を実行している。従って、全体的に最適化され

た単語分類結果を得ることができる。これにより、テキストデータから単語の分類体系を自動的に獲得するとき、より精密で正確な分類体系を得ることができる。

【0083】また、請求項2記載の単語分類処理方法は、請求項1記載の単語分類処理方法において、上記分類された複数 c 個のクラスを上記第2の記憶装置に記憶するステップは、上記第1の記憶装置に記憶された1つのウィンドウよりも外側のクラスが存在し、又は上記1つのウィンドウ内のクラスが c 個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行することの特徴とする。従って、所定の複数 c 個のクラスを有する中間層を最適化形式で得ることができる。

【0084】本発明に係る請求項3記載の単語分類処理装置は、複数の単語を含むテキストデータに対して、互いに異なるすべての複数 v 個の単語の出現頻度を調べ、出現頻度の高い単語から順に並べて、複数 v 個のクラスに割り当てる第1の制御手段と、上記複数 v 個のクラスの単語のうち出現頻度が高い v 個未満の $(c+1)$ 個のクラスの単語を1つのウィンドウ内のクラスの単語として第1の記憶装置に記憶する第2の制御手段と、上記第1の記憶装置に記憶された1つのウィンドウ内のクラスの単語に基づいて、互いに異なる第1のクラスの単語と第2のクラスの単語とが隣接して出現する確率を、上記第1のクラスの単語の出現確率と第2のクラスの単語の出現確率との積に対する相対的な頻度の割合を表わす所定の平均相互情報量が最大となるように、上記複数の単語を二分木の形式で複数 c 個のクラスに分類し、分類された複数 c 個のクラスを、単語分類結果を表わす全体のツリー図の中間層の複数 c 個のクラスとして第2の記憶装置に記憶する第3の制御手段と、上記第2の記憶装置に記憶された中間層の複数 c 個のクラスに基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまで分類し、当該分類結果を上記ツリー図の上側層として第3の記憶装置に記憶する第4の制御手段と、上記第2の記憶装置に記憶された中間層の複数 c 個のクラスの各クラス毎に、上記中間層の複数 c 個のクラスの各クラス内の複数の単語に基づいて、上記平均相互情報量が最大となるように、上記複数の単語を二分木の形式で1個のクラスになるまでそれぞれ分類し、当該各クラス毎の複数の分類結果を上記ツリー図の下側層として第4の記憶装置に記憶する第5の制御手段と、上記第4の記憶装置に記憶された上記ツリー図の下側層を、上記第2の記憶装置に記憶された上記中間層の複数 c 個のクラスと連結する一方、上記第3の記憶装置に記憶された上記ツリー図の上側層を、上記第2の記憶装置に記憶された上記中間層の複数 c 個のクラスと連結することにより、上側層と中間層と下側層とを備えた上記ツリー図を求めて単語分類結果として

第5の記憶装置に記憶する第6の制御手段とを備える。従って、下側層、中間層、及び上側層と階層化して、複数の単語を二分木の形式で複数のクラスに分類したので、単語分類処理によりバランスのとれた階層構造を有する単語分類結果を得ることができる。また、平均相互情報量の計算においては、下側層、中間層、及び上側層ともに、すべてのクラスの単語を対象として平均相互情報量を計算しているので、計算された平均相互情報量は局所的な平均相互情報量ではなく、全体の単語の情報を含んだグローバル平均相互情報量に基づいて、クラスタリング処理を実行している。従って、全体的に最適化された単語分類結果を得ることができる。これにより、テキストデータから単語の分類体系を自動的に獲得するときに、より精密で正確な分類体系を得ることができる。

【0085】また、請求項4記載の単語分類処理装置は、請求項3記載の単語分類処理装置において、上記第3の制御手段は、上記第1の記憶装置に記憶された1つのウィンドウよりも外側のクラスが存在し、又は上記1つのウィンドウ内のクラスがc個ではないときは、現在のウィンドウよりも外側にあり、最大の出現頻度を有するクラスの単語を上記ウィンドウ内に挿入した後、上記二分木の形式の単語分類処理を実行する。従って、所定の複数のc個のクラスを有する中間層を最適化形式で得ることができる。

【0086】本発明に係る請求項5記載の音声認識装置によれば、入力される発声音声の音声信号に基づいて、請求項3又は4記載の単語分類処理装置によって複数の単語が複数のクラスに分類された単語分類結果を含む単語辞書と、所定の隠れマルコフモデルとを参照して上記発声音声を音声認識する音声認識手段を備える。従って、上記単語分類処理装置により得られた、バランスのとれた階層構造を有しかつ全体的に最適化された単語辞書に基づいて音声認識することにより、従来例に比較して高い認識率で音声認識することができる。

【図面の簡単な説明】

【図1】 本発明に係る第1の実施形態である音声認識装置のブロック図である。

【図2】 本発明に係る第2の実施形態である形態素及び構文解析装置のブロック図である。

【図3】 図1及び図2の単語分類処理部によって実行される単語分類処理における加算領域及び加減算処理を示すクラスバイグラム平面テーブルの図である。

【図4】 (a)は上記単語分類処理におけるAMI減少量 $L_k(l, m)$ に対する加算領域を示すクラスバイグラム平面テーブルの図であり、(b)は上記単語分類処理におけるAMI減少量 $L_k^{(i,j)}(l, m)$ に対する加算領域を示すクラスバイグラム平面テーブルの図である。

【図5】 (a)は上記単語分類処理におけるマージ後

のAMI量 $I h_k$ を示すクラスバイグラム平面テーブルの図であり、(b)は上記単語分類処理におけるマージ後のAMI量 $I h_k(l, m)$ を示すクラスバイグラム平面テーブルの図であり、(c)は上記単語分類処理におけるマージ後のAMI量 $I h_{k-1}^{(i,j)}$ を示すクラスバイグラム平面テーブルの図であり、(d)は上記単語分類処理におけるマージ後のAMI量 $I h_{k-1}^{(i,j)}(l, m)$ を示すクラスバイグラム平面テーブルの図である。

【図6】 上記単語分類処理によって得られるデンドログラム(ツリーの系統図)の一例を示す図である。

【図7】 上記単語分類処理によって得られる左側方向の分岐ツリーの一例を示す図である。

【図8】 上記単語分類処理によって得られる1つのクラスに対するサブツリーの一例を示す図である。

【図9】 本発明の音声認識装置におけるシミュレーション結果である、テキストの大きさに対するパープレキシティーを示すグラフである。

【図10】 図1の単語分類処理部20の構成を示すブロック図である。

【図11】 図1の単語分類処理部20によって実行されるメインルーチンの単語分類処理を示すフローチャートである。

【図12】 図11のサブルーチンの初期化処理(S1)を示すフローチャートである。

【図13】 図11のサブルーチンの中間層クラスタリング処理(S2)を示すフローチャートである。

【図14】 図11のサブルーチンの上側層クラスタリング処理(S3)を示すフローチャートである。

【図15】 図11のサブルーチンの下側層クラスタリング処理(S4)を示すフローチャートである。

【図16】 図11のサブルーチンのデータ出力処理(S5)を示すフローチャートである。

【図17】 図11のサブルーチンの中間層クラスタリング処理(S2)におけるステップS21の処理を示し、単語クラスの集合を示す図である。

【図18】 図11のサブルーチンの中間層クラスタリング処理(S2)におけるステップS23及びS24の処理を示し、単語クラスの集合を示す図である。

【図19】 図11の単語分類処理における処理及びその処理によって得られる階層構造を示す図である。

【図20】 本発明の音声認識装置のシミュレーション結果である、テキストの大きさに対するクラスタリング処理後のラベル付けの誤り率を示すグラフである。

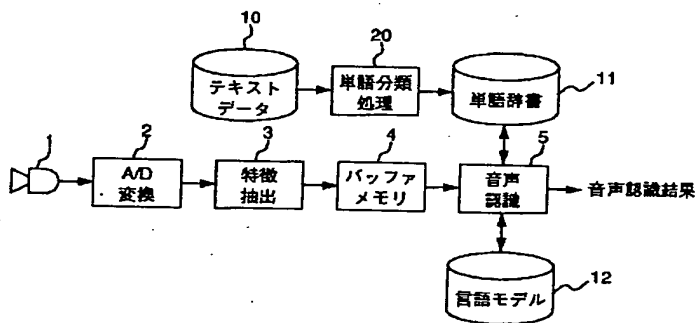
【符号の説明】

- 1…マイクロホン、
- 2…A/D変換器、
- 3…特徴抽出部、
- 4…バッファメモリ、
- 5…音声認識部、
- 10, 31, 32…テキストデータメモリ、

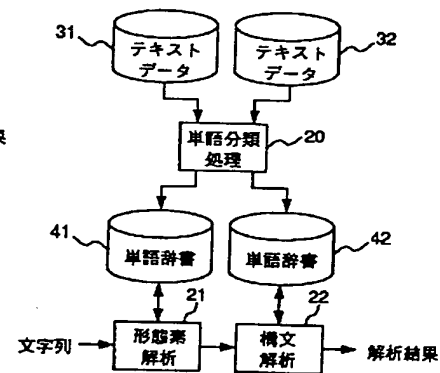
11, 41, 42…単語辞書メモリ、
 12…言語モデル、
 20…単語分類処理部、
 21…形態素解析部、
 22…構文解析部、
 50…CPU、
 51…ROM、
 52…ワークRAM、
 53…RAM、
 54, 55…メモリインターフェース、
 61…初期化クラス単語メモリ、
 62…AMIメモリ、
 63…中間層メモリ、

64…上側層ヒストリメモリ、
 65…上側層ツリーメモリ、
 66…下側層ヒストリメモリ、
 67…下側層ツリーメモリ、
 68…ツリーメモリ、
 100…中間層、
 101…上側層、
 102…下側層、
 S1…初期化处理、
 S2…中間層クラスタリング処理、
 S3…上側層クラスタリング処理、
 S4…下側層クラスタリング処理、
 S5…データ出力処理。

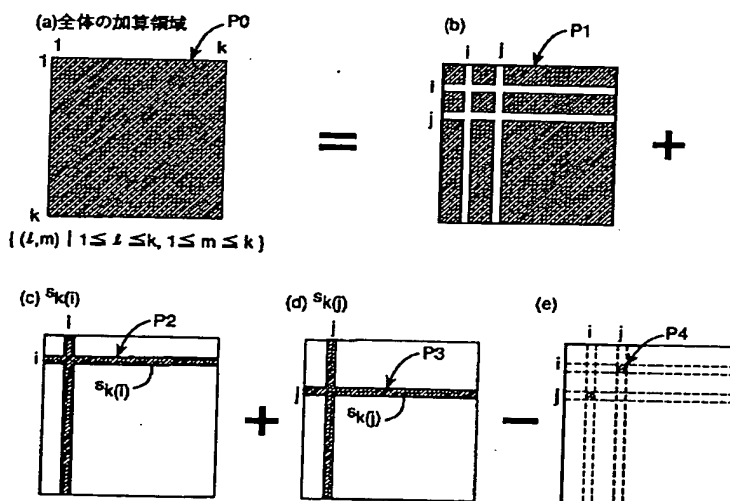
【図1】



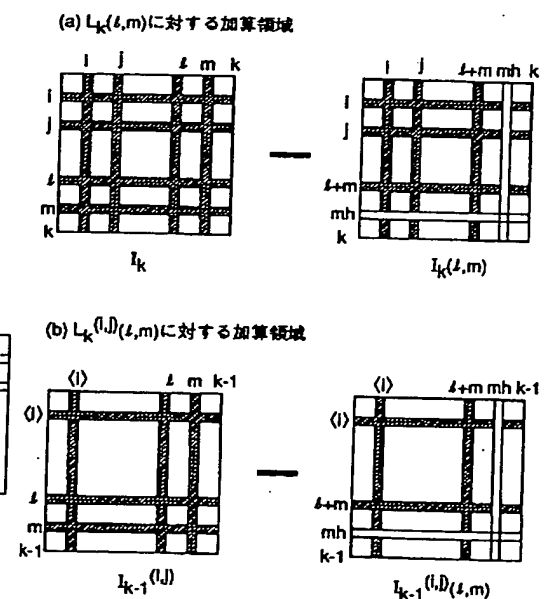
【図2】



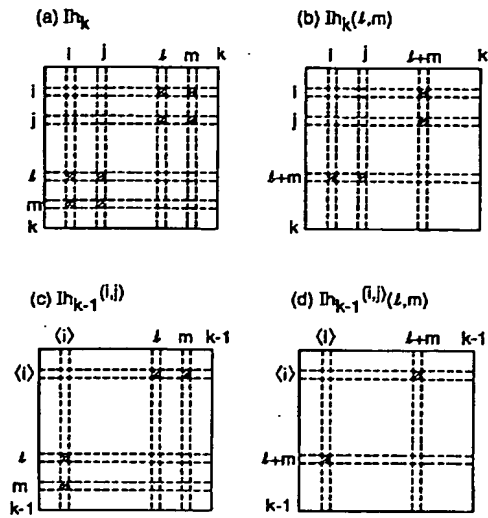
【図3】



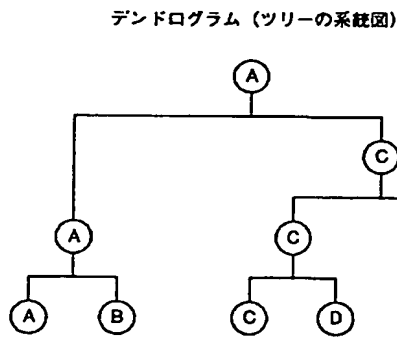
【図4】



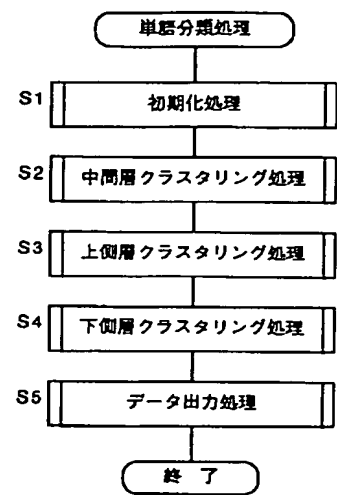
【図5】



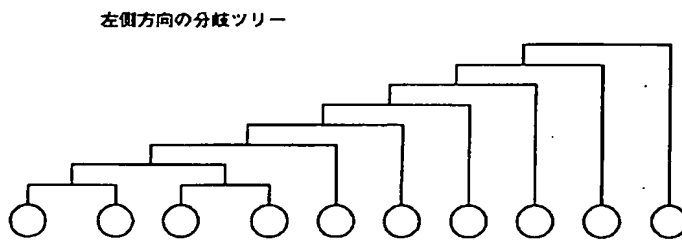
【図6】



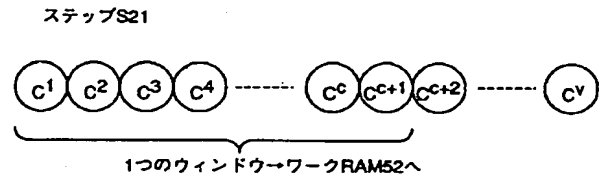
【図11】



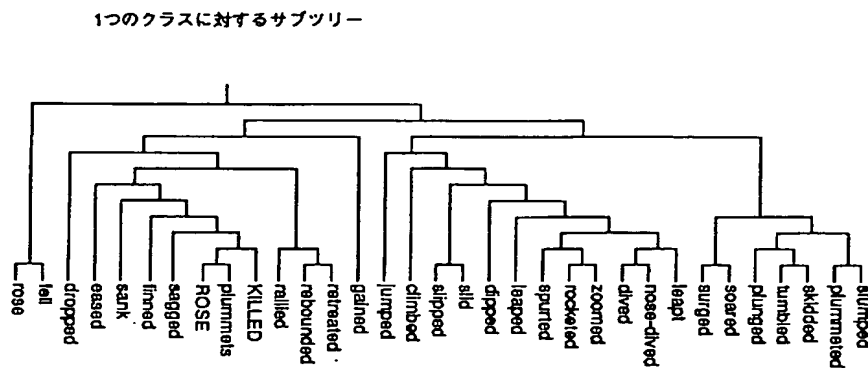
【図7】



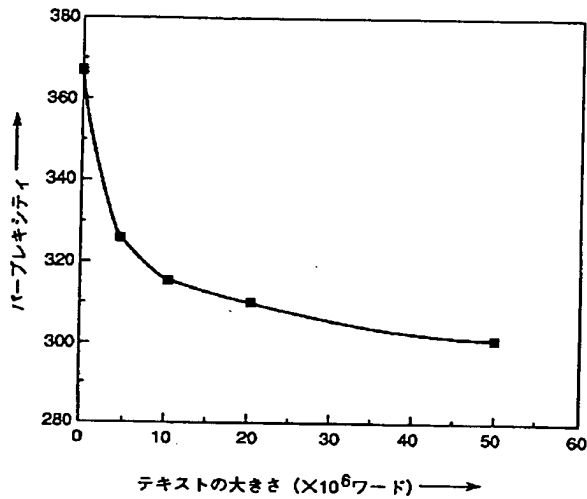
【図17】



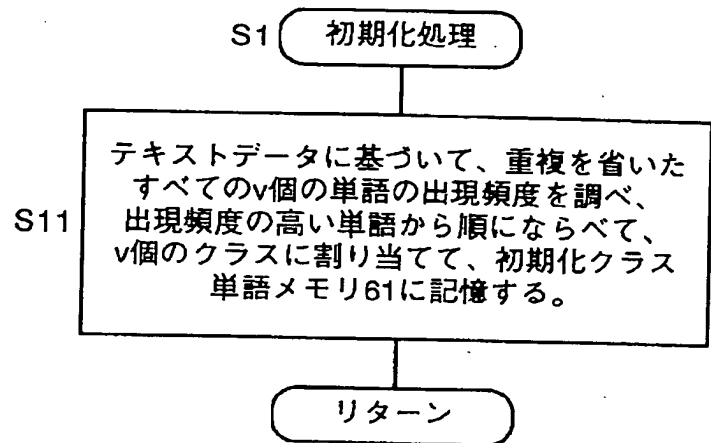
【図8】



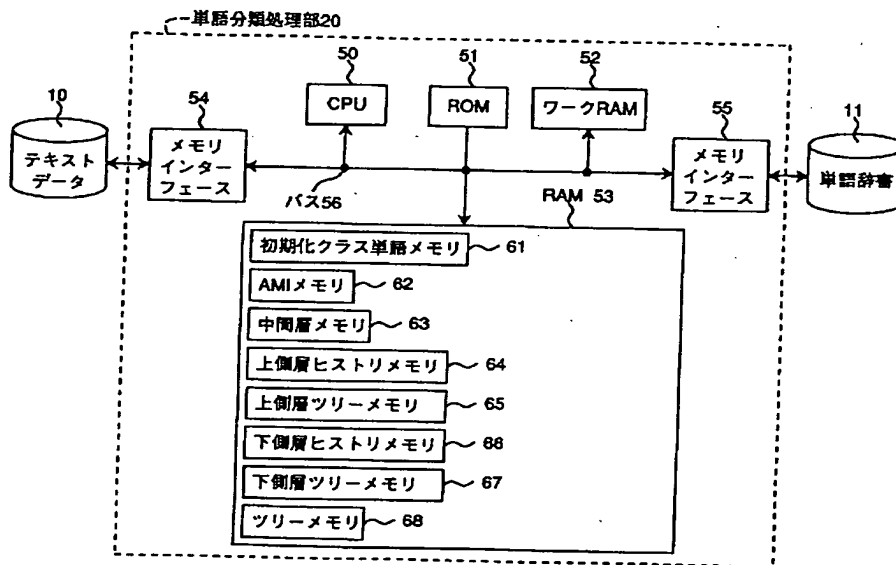
【図9】



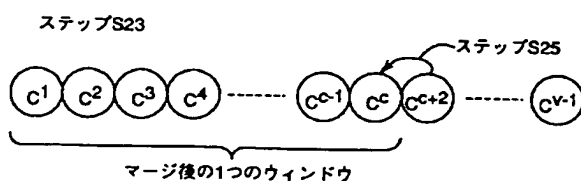
【図12】



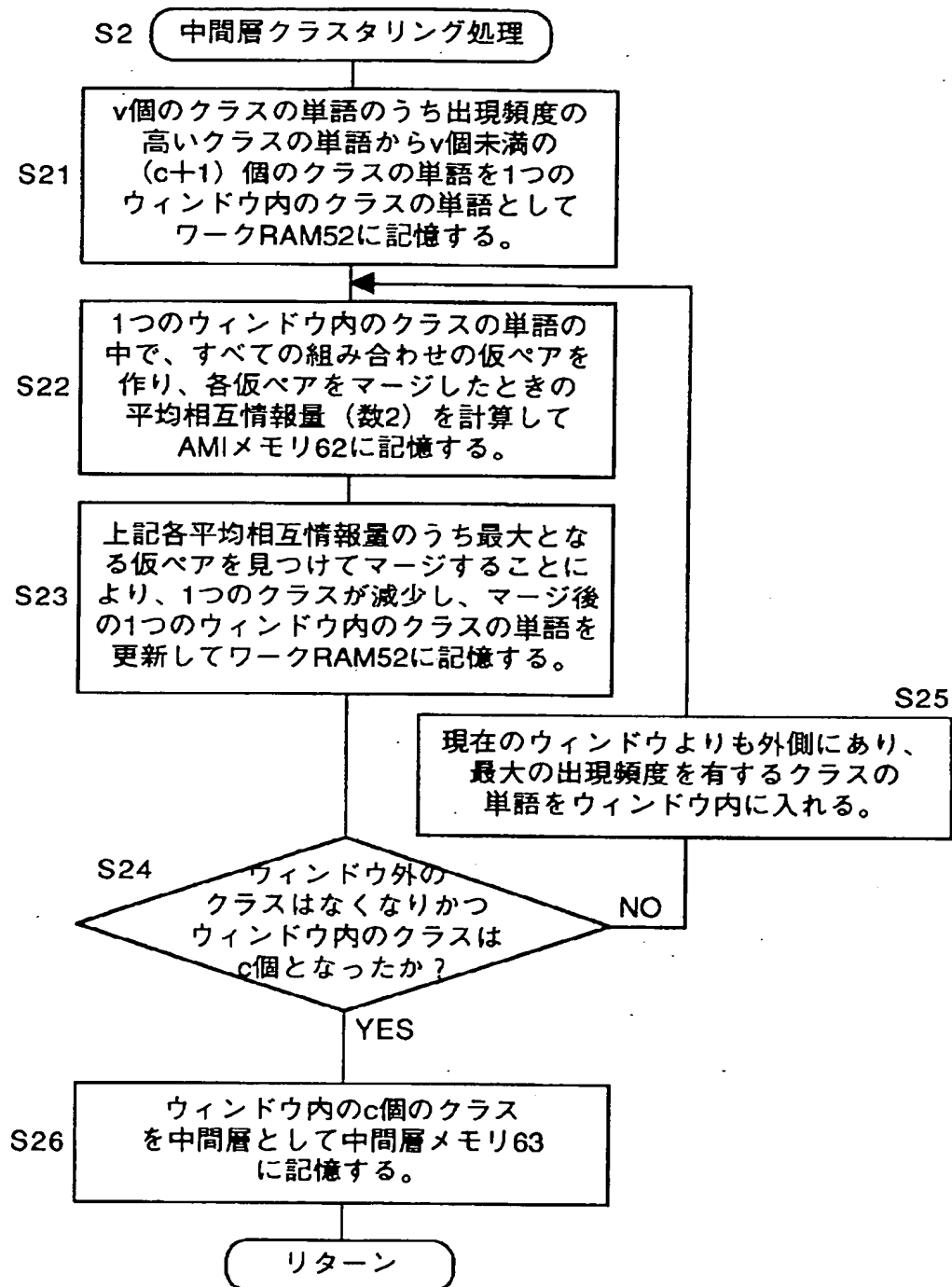
【図10】



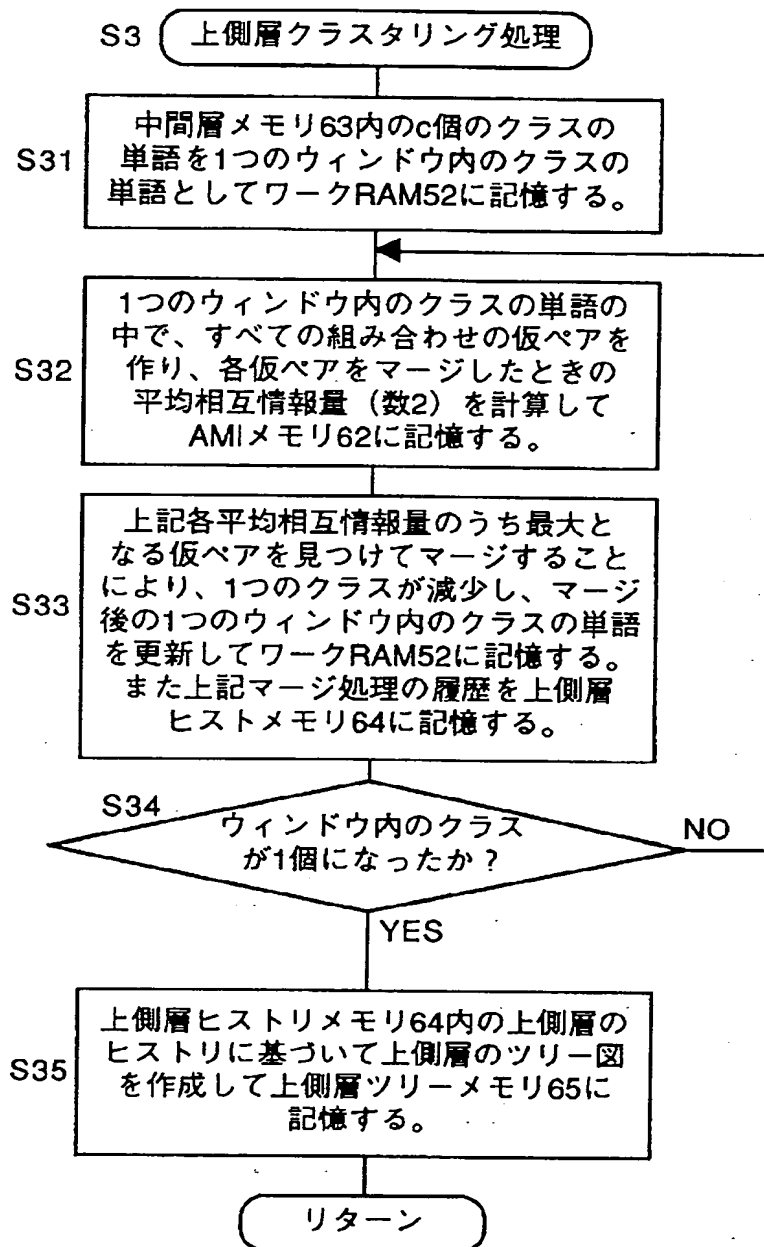
【図18】



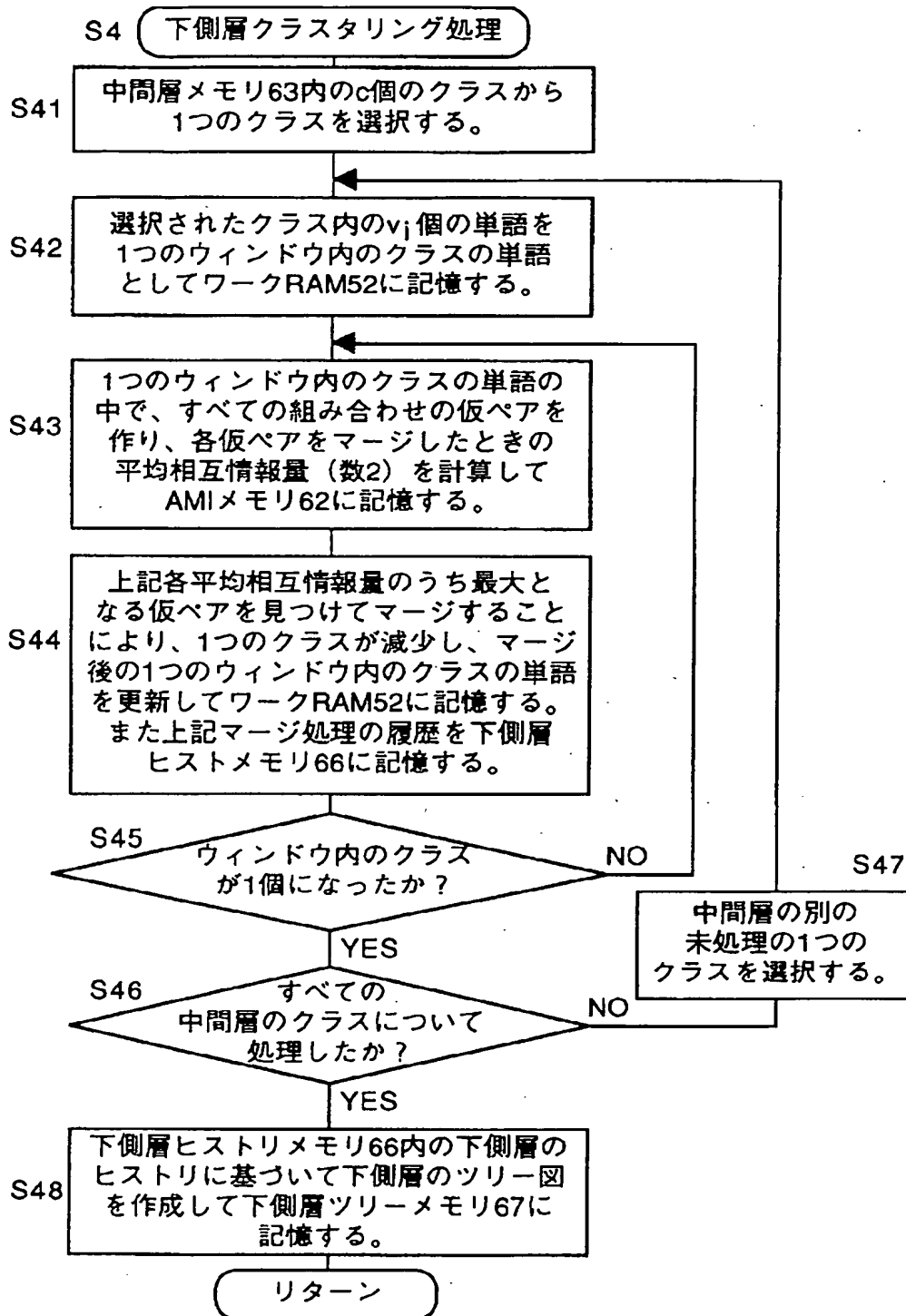
【図13】



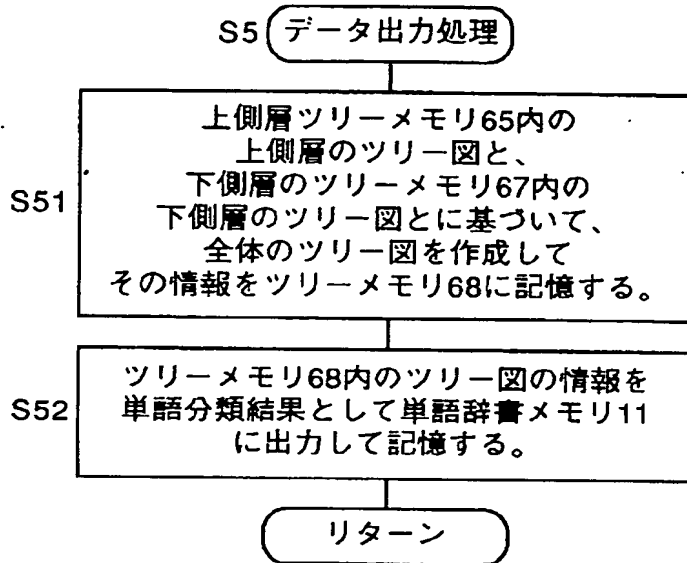
【図14】



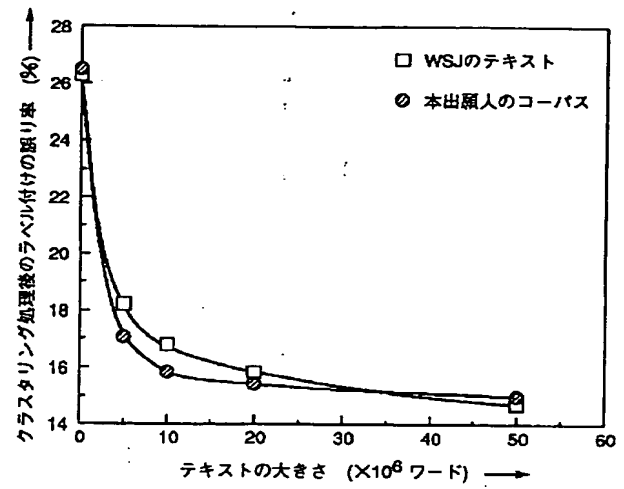
【図15】



【図16】



【図20】



【図19】

